Gender bias in machine learning for sentiment analysis[1]

Mike Thelwall
Statistical Cybermetrics Research Group, University of Wolverhampton, UK.

Purpose: This paper investigates whether machine learning induces gender biases in the sense of results that are more accurate for male authors than for female authors. It also investigates whether training separate male and female variants could improve the accuracy of machine learning for sentiment analysis.
Design/methodology/approach: This article uses ratings-balanced sets of reviews of restaurants and hotels (3 sets) to train algorithms with and without gender selection.
Findings: Accuracy is higher on female-authored reviews than on male-authored reviews for all data sets, so applications of sentiment analysis using mixed gender datasets will over represent the opinions of women. Training on same gender data improves performance less than having additional data from both genders.
Practical implications: End users of sentiment analysis should be aware that its small gender biases can affect the conclusions drawn from it and apply correction factors when necessary. Users of systems that incorporate sentiment analysis should be aware that performance will vary by author gender. Developers do not need to create gender-specific algorithms unless they have more training data than their system can cope with.
Originality/value: This is the first demonstration of gender bias in machine learning sentiment analysis.
Keywords: Sentiment analysis, opinion mining, gender, gender bias.

Introduction

Automatic sentiment analysis algorithms (Liu, 2012; Pang & Lee, 2008) are part of the standard toolkit for market research and customer relations management (Lamont, 2014; Liyakasa, 2012). They can detect the opinions of customers unobtrusively on a large scale in near real-time from their social web posts. Sentiment analysis is also used in government and politics to detect public opinion about issues or candidates (Wood, 2016; Wright, 2015). Analysts may implicitly assume that sentiment analysis results are unbiased because they are automatic but this is not necessarily true. Given the existence of clear gender differences in communication styles on the social web (Burger, Henderson, Kim, & Zarrella, 2011; Mihalcea & Garimella, 2016; Volkova & Yoram, 2015), including for expressing sentiment (Montero, Munezero, & Kakkonen, 2014; Thelwall, Wilkinson, & Uppal, 2010), interpreting sentiment (Guerini, Gatti, & Turchi, 2013) and discussing products (Yang, Kotov, Mohan, & Lu, 2015), gender biases in sentiment analysis seem likely. In other words, sentiment analysis algorithms may be more able to detect sentiment from one gender than from another so that, in a gender-mixed collection of texts, sentiment analysis results could over represent the opinions of one gender.

Thus, for example, a company exploiting sentiment analysis for customer relations management (Waller & Fawcett, 2013) might find that 50% of female-authored reviews of their product were positive in comparison to 40% of male-authored reviews. From this they might conclude that their product is more appealing to females, whereas the 10% difference

[1] Thelwall, M. (in press). Gender bias in machine learning for sentiment analysis. Online Information Review. doi: 10.1108/OIR-05-2017-0139

might be due to the algorithms being better able to detect sentiment expressed by females. Similarly, a company with two products, A and B, where B is an upgraded variant of A might find that A has a 50% approval rating whilst B has a 60% approval rating. If females preferred B but males preferred A and the sentiment analysis algorithm detected female sentiment better, then it is possible that A and B have the same overall approval rating but that gender bias in the algorithm (i.e., greater ability to detect sentiment from females) caused the difference in the overall ratings. Gender bias may also affect the performance of systems that detect sentiment to deliver personalised experiences, such as in the healthcare sector (Hadžidedić Baždarević, & Cristea, 2017). Although gender bias has been found for lexical algorithms (Thelwall, in press), it is unclear whether it is also present in machine learning.

Is it already known that machine learning algorithms can learn society's gender prejudices and then reinforce them. For example, a core component of some algorithms is the ability to deduce the meaning of words by associating them with other words that tend to occur in the same document. Using this approach can lead to conservative implications, such as that *homemaker* is part of the "meaning" of the word *woman* and that *programmer* is part of the meaning of the term *man* (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). This could lead to adverts for programmers to be delivered mainly to male users, targeting women with home appliance offers instead. Thus, algorithms can deliver misguided information, leading to bias in system outputs (such as adverts or search results) or the conclusions drawn by a human from these outputs (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016).

For machine learning-based sentiment analysis, a previous study has developed an improved sentiment classifier that takes gender into account by ignoring features (words or phrases) that have different polarity associations for males than for females (Volkova, Wilson, & Yarowsky, 2013). This shows that gender information can help machine learning even when using a single classifier. Another study optimised lexical classifiers on monogender collections of reviews (Thelwall, in press), without finding statistically significant improvements from this approach. This lack of improvement could be due to the knowledge-rich approach of lexical sentiment analysis masking the relatively small gender differences. Alternatively, the main gender differences may occur in words or phrases that are not in the sentiment lexicon because they are not explicitly sentiment-bearing. A study of review author attributes in Chinese-language Sina Weibo posts found that including gender as a feature can improve polarity detection (Li, Yang, & Zong, 2016). Despite these papers, no previous study has assessed whether gender bias is present in machine learning classifiers or, as a secondary issue, whether gender specific machine learning classifiers can be more accurate than gender neutral classifiers.

Background: Algorithmic bias

There is a growing concern that data processing algorithms are influencing policy and daily lives in ways that are not transparent and introduce biases. Key complex data processing tasks that could incorporate bias include *prioritisation* of information when there is too much for human consumption (e.g., in search engine results), *classification* for attributes of texts, images or other objects (e.g., positive or negative opinions), *association* of related objects, and *filtering* out objects deemed spam or irrelevant (Diakopoulos, 2015). Algorithms that perform these tasks are frequently opaque due to complexity or commercial secrecy. End users then have little chance to check for biases in the results that

they receive, especially if it is an unsolicited technology, such as advertising. The following examples from news provision, search engine results, advertising, and academia illustrate the potential impacts of algorithmic bias.

News agencies use programs to automatically generate summaries from data feeds. These summaries are then used by news media for their stories. Similarly, Facebook uses algorithms to select content for users' newsfeeds (Diakopoulos, 2016). In both cases, users risk having their view of the world influenced by biases or errors in the algorithms, whether intended (e.g., for commercial gain) or accidental. Deliberate or accidental news algorithm biases may even untraceably affect election results (Tufekci, 2015). There is a long tradition of press bias through selecting which stories to present (Herman & Chomsky, 2010) but algorithmic biases are more insidious by originating from an apparently impartial source that cannot be easily challenged, especially when they are a component part of a larger system (e.g., Facebook) or part of a pipeline (e.g., news summary generation by news agencies).

Search engines are increasingly the main portal for accessing all types of information. Although the major search engines do not allow URLs to be sponsored in their main results lists, they use complex algorithms to rank URLs and select those to show (Halavais, 2013). This gives international biases (Vaughan, & Thelwall, 2004) and can reinforce societal inequalities. For example, image search results for science may return photographs dominated by males (Kay, Matuszek, & Munson, 2015). Bias can also appear as part of personalisation, when properties inferred about the user are exploited to decide which URLs to return in their search results pages (Teevan, Dumais, & Horvitz, 2004). This could deliver URLs deemed appropriate based on the average wealth, ethnicity or age of the typical occupants of a location, and this partial information segregation can be socially undesirable.

Modern online targeted advert placement software uses heuristics to estimate the most receptive users for a campaign and can exacerbate pre-existing societal inequalities. It can tend to deliver high-paying job adverts to males (Datta, Tschantz, & Datta, 2015) or negative adverts (e.g., for prison record searches) in response to ethnic minority name searches (Sweeney, 2013).

Academic social network sites like ResearchGate.net prominently display summary information about researchers on their home pages without explaining where these are from (Jordan, 2015). Since these sites may be used for hiring and reputation management, they may be unaccountably influencing whether individuals have a successful career, and perhaps even the relative success of different nationalities (Thelwall & Kousha, 2015). An example of good practice in this context is Altmetric.com, which tracks mentions or citations of academic research online and clearly delineates its primary data, which has a verifiable audit trail, from the remaining data (Adie & Roe, 2013; Liu & Adie, 2013; for an overview, see: Williams, 2017).

Issues like those above have led to calls for algorithmic accountability and transparency in the sense of the right of society to check that key algorithms behave ethically (Diakopoulos, 2016). Transparency can be impractical due to the complexity of an algorithm – if few understand it – or unethical because it uses private information, such as search engine log files, to learn its behaviour (Ananny & Crawford, in press). Nevertheless, its seems intuitively to be a social good to detect biases in algorithms, when possible, and to compensate for them or take them into account. This issue is also being raised with software developers and is starting to be taken seriously in computing (e.g., de O Melo & de

Sousa, 2017; Hajian, Bonchi, & Castillo, 2016). There may also be legal implications in some cases. If it could be proven that a company had advertised online for hiring and the advert had been shown disproportionately to males, then this may be ruled to be illegal in some countries. Employers would then need to ensure that they only used demonstrably gender neutral algorithms.

An algorithm can also be biased by gender, ethnicity, nationality, wealth or other social characteristic if it performs differently or with different levels of efficiency for different groups. For example, facial recognition technology (Han, Otto, Liu, & Jain, 2015) needs careful construction to be able to recognise all ethnicities with a similar level of performance (Klare, Burge, Klontz, Bruegge, & Jain, 2012; the same is true for humans: Chiroro, Tredoux, Radaelli, & Meissner, 2008). This can lead, for example, to increased arrest rates amongst the ethnicities recognised most easily from Closed Circuit TV footage. Any learning algorithm is likely to most efficient for types of user that are most represented in its training dataset (see also: Hargittai, 2015). Hence, if a social network site has predominantly young, white, female American users then its news algorithm, if not personalised, is likely to select stories that interest young, white, female American users. More subtly, if the algorithm finds it easier to detect the behaviours of one group because they are more explicit or homogenous then it will tend to be overly influenced by this group. This is not only a societal problem for the use of such algorithms but also a commercial issue. For instance, if biased algorithms are used for customer relations management (e.g., Lamont, 2014; Liyakasa, 2012) then they may deliver misleading overall customer information or unfair comparisons between customer groups.

As mentioned in the introduction, social web communication styles vary on the social web in general (Burger, Henderson, Kim, & Zarrella, 2011; Mihalcea & Garimella, 2016; Volkova & Yoram, 2015), for expressing sentiment (Montero, Munezero, & Kakkonen, 2014; Thelwall, Wilkinson, & Uppal, 2010), and for evaluating products (Yang, Kotov, Mohan, & Lu, 2015). In many different (mainly offline) contests, males seem more inclined to discuss aspects of objects whereas females are more likely to refer to psychological and social issues (Newman, Groom, Handelman, & Pennebaker, 2008). Females are more likely to express sadness, anxiety and positive feelings, although men are more likely to swear, which can be indicative of emotion. A study of Facebook posts from 75,000 users found females to be more likely to discuss affective processes and to use words related to positive emotion, anxiety and sadness, whereas males expressed more negative emotions and anger (Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013). Thus, explicit sentiment and positivity seem likely to be more frequent in female-authored texts, although negativity and anger may be more common for males.

Research questions

The research questions address ways in which gender may influence the outcome of, or be used to improve, machine learning sentiment analysis. The primary issue of assessing whether gender bias is present in the results of machine learning for sentiment analysis is first. The remaining three research questions address different gender-based methods of improving the accuracy of machine learning sentiment analysis.

1. Is there a gender difference in the accuracy of machine learning sentiment analysis ratings predictions on review texts?
2. Is the accuracy of ratings predictions from male-authored review texts improved if the training corpus is exclusively male-authored?

3. Is the accuracy of ratings predictions from female-authored review texts improved if the training corpus is exclusively female-authored?
4. Are the advantages of single gender training, if any, greater than the disadvantage of reducing the size of the training set by omitting one gender author?

Methods

Data

Reviews were extracted from the TripAdvisor.com consumer reviews of hotels and restaurants. Since there are nationality differences in language and international differences in the aspects of products and services that are valued, it is important to generate homogenous datasets to seek fine grained differences. For this reason, only reviews written by UK residents about hotels and restaurants within England were selected for analysis. The TripAdvisor sitemap was downloaded and an apparently complete list of 13,234,039 URLs of reviewed entities in England was extracted. A random sample of 10% of these was downloaded between February and March 12, 2017 at a maximum rate of 1 URL per second. All reviews were extracted from all pages, together with the location and identity of the reviewer, giving a total of 6,121,296 reviews (there were multiple reviews on most pages). Reviewers that did not give a location within the UK were also discarded, leaving 3,561827 reviews. To qualify, users had to have an address naming the UK or one of its constituent countries, or one of the 20 largest UK cities. Entities that were not hotels or restaurants were discarded.

Reviews are accompanied by the reviewer's rating of the hotel or restaurant (10, 20, 30, 40 or 50) and these were also extracted. User ratings are appropriate proxies for sentiment strength (e.g., for sentiment polarity, see: Turney, 2002; Pang & Lee, 2004) because higher ratings indicate that the reviewer is more positive about the product.

Reviewer genders were detected from their first name or the first part of their username (e.g., Jane in "Jane Smith" or Jazzy in "JazzyG4532"). People with a first name of at least four letters and being at least 90% male or 90% female in the 1990 US census were classified with the matching gender or were otherwise discarded. This process is over 90% accurate, by design, and as verified by manual checking of the data, but rejects about 70% (depending on the sample) of the data due to short or ambiguous first names. The 1990 US census was used because more recent data is not available and no comparable UK data has been released. Cultures that are common in the UK are also represented in the US and so the US source is a reasonable substitute.

In addition, since there are wide differences in customer expectations of different types of hotel, the hotels were split by star grades. Due to the restricted amount of data, only two of the five main star grades were used: 3.0 (299,954 reviews), and 4.0 (281,091 reviews), and a merged set for all hotel reviews (720,897 reviews). Together with restaurants (571,569 reviews), this gave four core relatively homogenous datasets. Identical experiments were performed on all four datasets to reduce by replication the risk that the results are specific to a product type.

For each topic, this produced three data sets: one with equal numbers of male and female authors (MF), one with only male authors (M), and one with only female authors (F), each with equal numbers of texts with hotel/restaurant ratings of 10, 20, 30, 40 and 50. A maximum of one review was allowed per reviewer. The requirement for unique reviewers, author genders to be assigned (about 30% had a gender detected), single gender training

sets in some cases, and equal numbers of reviews for each rating level (see below – about 6% of texts had a rating of 10) reduced the effective sizes of the samples.

This study used only shared public texts and was exempt from ethical approval.

Experiments

The experiments were run for sample sizes of 1,000, 2,000, 4,000 and 8,000 reviews (when sufficient data was available), with equal numbers of texts at each of the five ratings levels. The following experiments were run, each with 30 repetitions of 10-fold cross validations to allow confidence intervals to be calculated.

1. Same gender training for males: Training on M, evaluation on M.
2. Cross gender training for males: Training on F evaluation on M.
3. Both gender training for males: Training on MF, evaluation on M.
4. Cross gender training for females: Training on M, evaluation on F.
5. Same gender training for females: Training on F evaluation on F.
6. Both gender training for females: Training on MF, evaluation on F.
7. Both gender training for both genders: Training on MF, evaluation on MF.

All training on males was on the same set of training files (for 1, 4), and the same for females (2 and 5) and both gender files (3, 6, 7). The evaluation texts were the standard left out 10% for training on the same gender mix dataset as the evaluation (1, 5, 7). For cross gender training the complete set of texts of the appropriate gender and same size corpus was used for evaluation, since there was no overlap (2, 4). For training on the both gender set and evaluation on a single gender set, a new single gender evaluation set was constructed for both genders to exclude texts within the MF set and keep a balanced collection. Thus, in all cases the evaluation did not include any texts within the training set and the evaluations for different training sets used the same evaluation sets as far as possible.

Two classifiers were used on the data in order to classify the texts for the ratings 10, 20, 30, 40, 50 (i.e., a five class problem): Naïve Bayes and Support Vector Machines (SVM), as implemented in Weka. These are differing types of classifier that are known to give good results in this type of task (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). Of these, SVM was substantially more accurate and so its output was reported. All word, punctuation or emoticon unigrams, bigrams or trigrams (e.g., *love*; *to the*; *very nice!!!*; *mice :(are*) were extracted from each text as its feature set. The restriction to 1-3 grams was made because the use of longer n-grams does not improve machine learning accuracy (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). Feature selection was accomplished using correlation as this performed much better than feature selection with information gain. A simple heuristic was used of choosing equal numbers of features with high positive and negative correlations. Correlations between features tended to be low and so, given the low number of features selected overall, grouping (Hall, 2000) was not used. An optimal feature size was identified for each corpus size by taking the optimal out of 100, 200, … 1000 features. Feature selection was performed only on the training sets in each case (so for each 10-fold cross-validation it was used 10 times).

Since 1000 texts is a relatively small number, additional steps were taken to get better evaluation results for this corpus size. For this, a dataset was built of 4000 texts and then the machine learning accuracy was assessed by splitting the set into 10 random folds where the training set was 1000 texts and the evaluation set was the remaining 3000 texts. This is more powerful than standard 10-fold evaluation, the training set varies over 4000

texts rather than 1000 and the evaluation set is larger each time and includes 4000 texts in total. Thus, the results from this experiment are less sensitive to the specific choice of texts.

Results

The results are discussed separately for each research question.

RQ1: Gender difference in machine learning accuracy

For almost all topics, training set compositions and data set sizes, female-authored reviews are more accurately classified than male-authored reviews (RQ1: compare the ->M bars with the ->F bars in Figure 1 and the same for Figures 2, 3 and 4). Results should only be compared within individual figures and not between different figures.

To explain the above claim, in Figure 1, the sentiment algorithm trained on the 4k bigender 3* hotels texts set but evaluated on (i.e., applied to) female-authored texts (3* hotels, 4k, MF->**F**) gives a correlation of about 0.78 with the human ratings. In contrast, when the algorithm is applied to male-authored texts (3* hotels, 4k, MF->**M**), the correlation is lower at 0.76. Similarly, for 3* hotels with the 4k training sets, when the algorithm is applied to female authored texts (3* hotels, 4k, F->**F**; 3* hotels, 4k, M->**F**; 3* hotels, 4k, MF->**F**) the correlations with human ratings are in the range 0.775-0.78, whereas for males the corresponding (3* hotels, 4k, F->**M**; 3* hotels, 4k, M->**M**; 3* hotels, 4k, MF->**M**) correlations are in the lower range 0.745-0.76. Thus, all three correlations for male-authored texts are lower than all three correlations for female-authored texts for the 4k 3* hotels set.

The above paragraph is also true (except for the exact value of the correlation figures) for the 2k and 1k data sets for 3* hotels (orange and blue bars in Figure 1). Thus, all three correlations for male-authored texts are lower than all three correlations for female-authored texts for the 2k 3* hotels set. Similarly, Thus, all three correlations for male-authored texts are lower than all three correlations for female-authored texts for the 1k 3* hotels set.

The above two paragraphs are also true for the 4* hotel reviews (Figure 2), all hotel reviews (Figure 3) and for restaurants (Figure 4), including for the 8k data set. The only exception is that in the 8k set (Figure 4), the M->F bar is slightly shorter than the MF->M bar for All hotels. Thus, almost irrespective of the size or composition of the training set, the algorithm performs better on female-authored texts.

Thus sentiment (in the form of ratings) is intrinsically easier to detect in female-authored texts (using machine learning), irrespective of training set size and review type (with one exception). This gives a clear answer to the first (and main) research question.

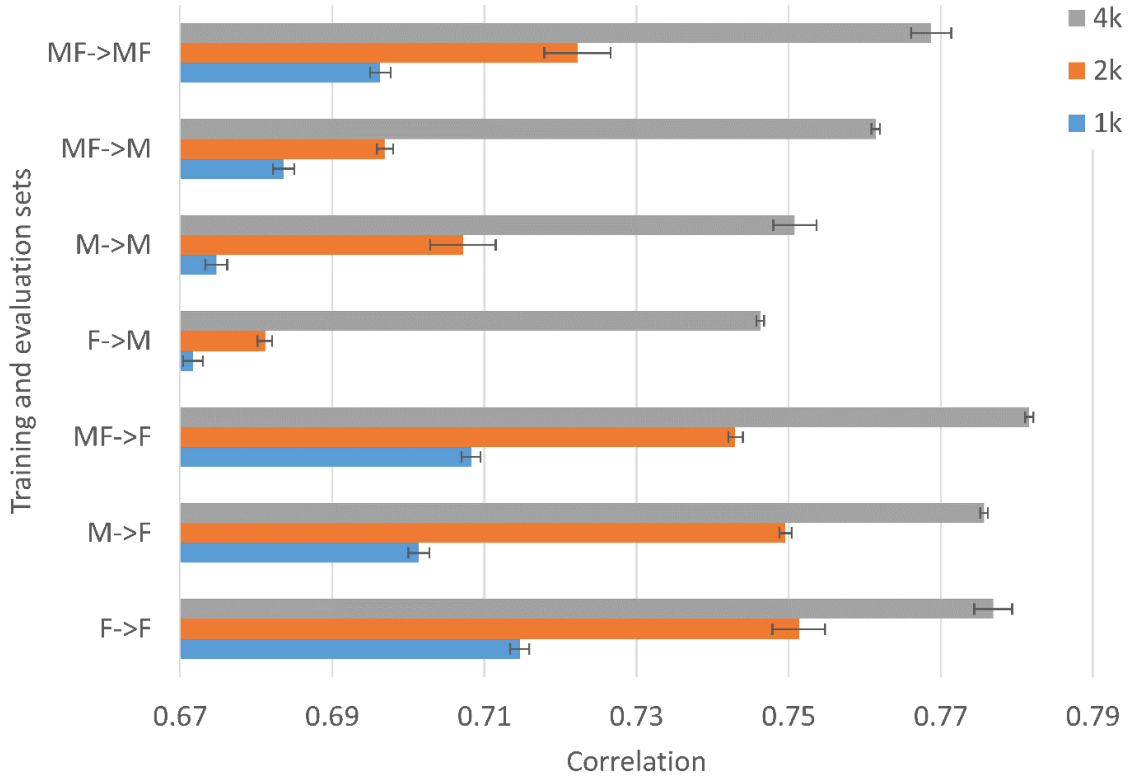

Figure 1. Correlations between SVM estimates and author ratings of TripAdvisor reviews for 3* hotels on different size rating-balanced corpora.

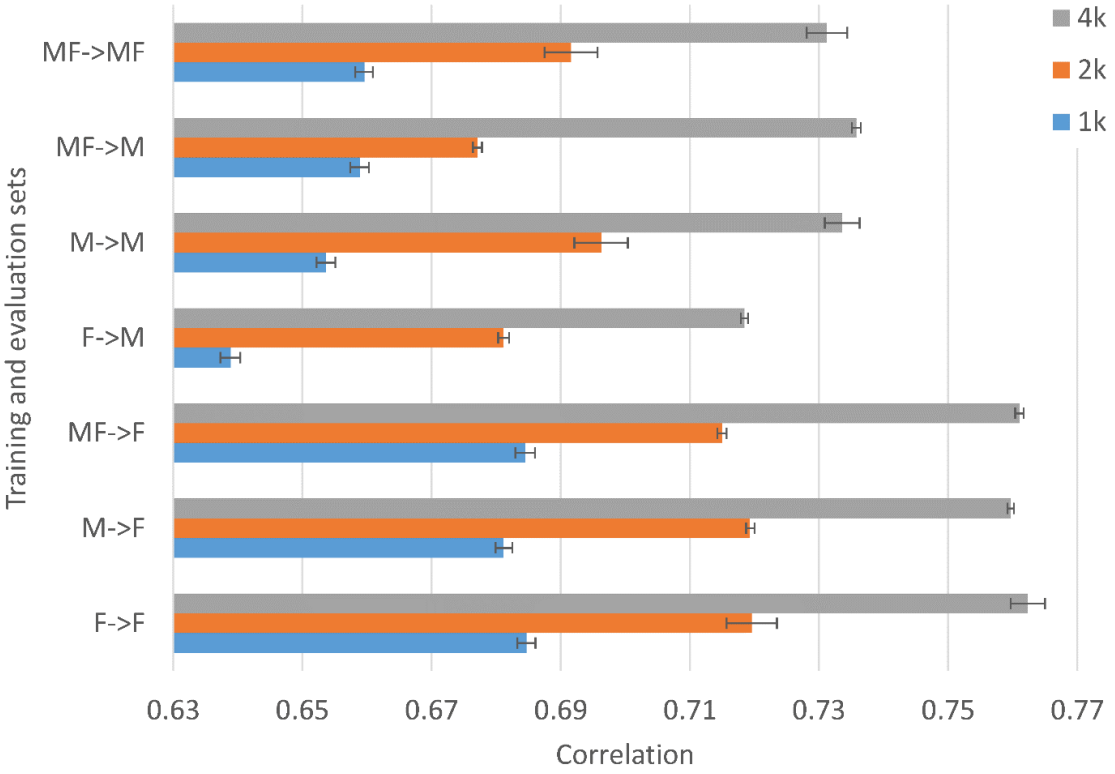

Figure 2. Correlations between SVM estimates and author ratings of TripAdvisor reviews for 4* hotels on different size rating-balanced corpora.

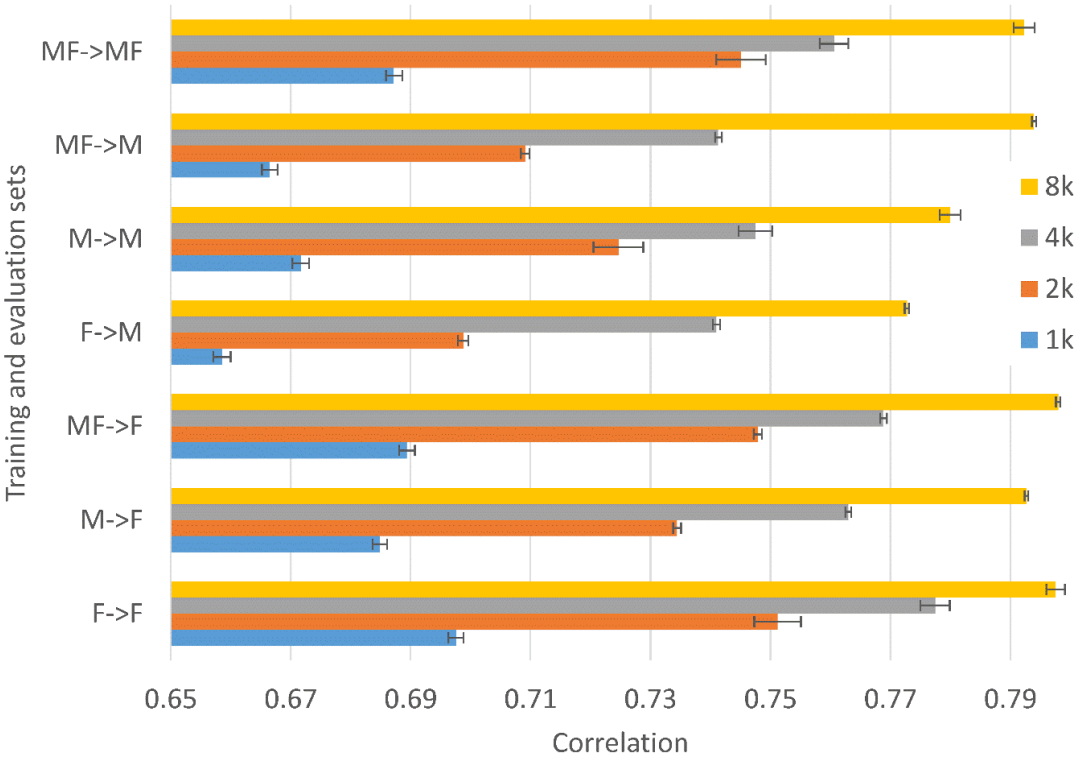

Figure 3. Correlations between SVM estimates and author ratings of TripAdvisor reviews for all hotels on different size rating-balanced corpora.

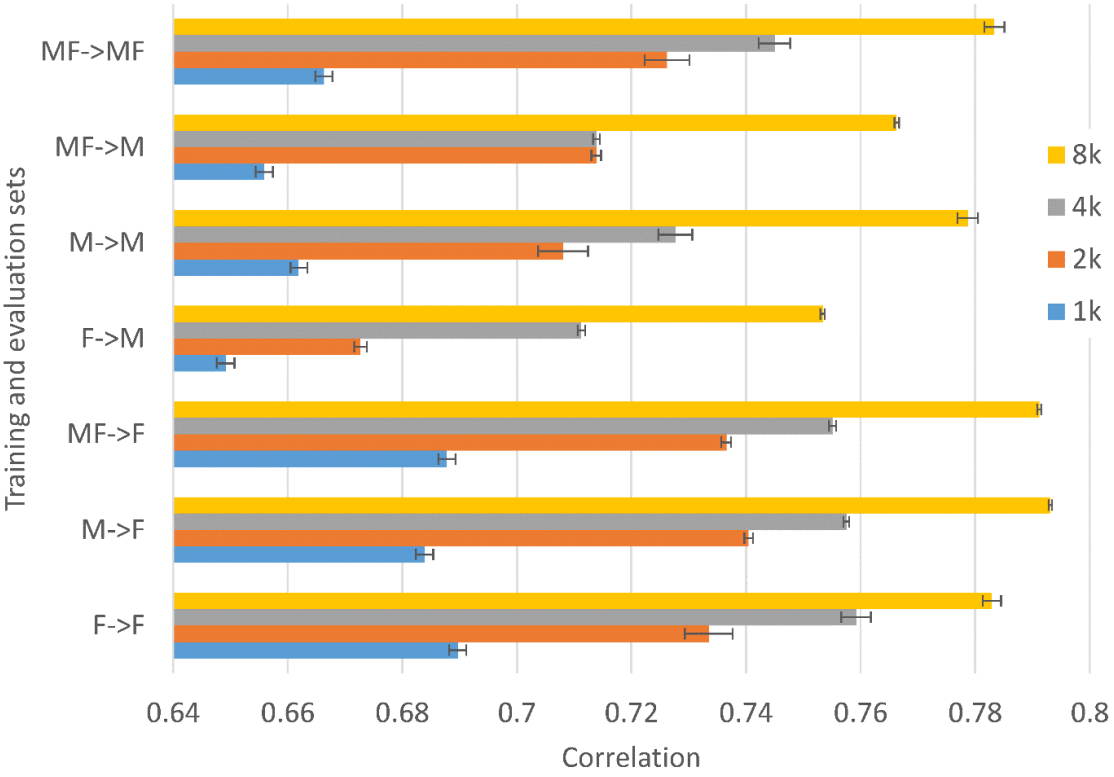

Figure 4. Correlations between SVM estimates and author ratings of TripAdvisor reviews for Restaurants on different size rating-balanced corpora.

RQ2: The accuracy of ratings predictions from male-authored review texts with a male-authored training corpus

For the second research question, the accuracy of ratings estimates for male-authored review texts tends to be highest if the training corpus is exclusively male-authored (i.e., M->M has a higher correlation than F->M or MF->M for the same training set size) [1000 texts: (2 out of 4): Restaurants, All hotels; 2000 texts (3 out of 4): All hotels, 3* hotels, 4* hotels; 4000 texts (2 out of 4): Restaurants, All hotels; 8000 texts (0 out of 2)], and lowest for female authored review texts [1000 texts (4 out of 4); 2000 texts (3 out of 4): Restaurants, All hotels, 3* hotels; 4000 texts (4 out of 4); 8000 texts (2 out of 2)]. Thus, training on male-authored review texts tends to be the optimal strategy for an algorithm designed to process male-authored review texts, if the training set size is fixed.

RQ3: The accuracy of ratings predictions from female-authored review texts with a female-authored training corpus

To answer the third research question, the accuracy of ratings predictions for female-authored review texts tends to be highest if the training corpus is female-authored (i.e., F->F has a higher correlation than M->F or MF->F for the same training set size) [1000 texts (4 out of 4); 2000 texts (3 out of 4): All hotels, 3* hotels, 4* hotels; 4000 texts (3 out of 4): Restaurants, All hotels, 4* hotels; 8000 texts (0 out of 2)], lowest for male authored review texts [1000 texts (4 out of 4); 2000 texts (1 out of 4): All hotels; 4000 texts (3 out of 4): All hotels, 3* hotels, 4* hotels; 8000 texts (1 out of 2): All hotels] and in between for mixed review texts. Thus, training on female-authored review texts tends to be the optimal strategy for an algorithm designed to process female-authored review texts, if the training set size is fixed.

RQ4: Single gender training vs. larger dual gender training

Comparing the performances between the different corpora other to answer the final research question (Figure 1-4), classification accuracy increases with corpus size more than the improvement for monogender training in all cases. For example, male training on 1000 texts is always less accurate than bigender training on 2000 texts. Thus, the loss of training data does not compensate for the increased relevance of monogender training.

Discussion and conclusions

The results are limited by using only two machine learning algorithms (the more accurate of which is reported above), one type of data (TripAdvisor reviews) from one nationality of reviewer (British) and the parsing options (e.g., not using part of speech tagging), the feature selection method (correlation), the feature scoring method (frequencies rather than alternatives such as TF-IDF: Paltoglou & Thelwall, 2010) and the corpus sizes. The automatic gender detection method introduces unknown biases, such as towards older users if younger users tend to use pseudonyms (e.g., flyboy33), and towards cultures with names that are common and gendered in the US (e.g., excluding many Sikhs with gender neutral first names and UK Malaysians using their given name as a second name). The results also only relate to one type of sentiment analysis task: predicting rating scores, rather than subjectivity, polarity (Turney, 2002) or fine grained emotion detection (Neviarouskaya, Prendinger, & Ishizuka, 2009). Thus, whilst the results demonstrate that sentiment analysis

machine learning can be gender biased, in the sense of being more accurate for one gender than another, they do not prove that it always or usually will.

The same limitations apply to the finding that machine learning is not able to improve accuracy by training on a same gender dataset because the reduction in training corpus size offsets any monogender improvement. Thus, the previously demonstrated strategy of removing gender-sensitive features (Volkova, Wilson, & Yarowsky, 2013) seems to be the best way of exploiting gender information to improve accuracy can also improve accuracy overall. This finding agrees with similar evidence that gender-specific lexical sentiment analysis is not more accurate (Thelwall, in press).

The lack of a consistent rank order for the correlations from the seven experiments (MF->MF; MF->F; MF->M; M->F; M->M; F->F; F->M) between the different datasets and corpus sizes in the results suggests that the role of gender for sentiment is not constant but is sensitive to the nature of the product being reviewed. In the above results, there were differences between hotels with different ratings. Whilst both 3* and 4* hotels are similar in broad function, their ratings would create greatly different expectations for their customers and so are distinct products in the same way that laptops and PCs are both computers but not all attributes that are important for one are also important for the other.

In summary, the experiments give clear evidence that machine learning-based sentiment analysis algorithms are better able to detect the sentiments of females than of males. They can sometimes perform better if trained separately on male and female texts, but monogender training is less effective overall because of the lower amount of training data. It should only be considered when there is more training data than a system can cope with. End users should be aware that their sentiment analysis results may slightly over-represent the opinions of females because a greater proportion of male sentiments will remain undetected. This is important in all contexts where gender is likely to influence sentiment. Similarly, if the opinions or males are compared to those of females, then small differences in the results could be due to algorithmic gender biases. In both cases, calibrating sentiment analysis accuracy separately for each gender would allow the results to be gender-corrected, such as by multiplying the results by males by a correction factor that would compensate for the greater number of undetected sentiments.

More generally, the findings add to the evidence of the ubiquity of algorithmic biases by showing that the results may be more accurate for one gender even when using an objective and impartial dataset and algorithm. This adds weight to current calls for algorithmic transparency. Because in this new case the bias is only in the output and not the inputs (data, algorithm), it is important that demands for algorithmic accountability should include the ability to test the outputs of a system, including the ability to drill down into different user groups by gender, ethnicity and other characteristics, to identify, and hopefully suggest corrections for, system output biases.

References

Adie, E., & Roe, W. (2013). Altmetric: enriching scholarly content with article-level discussion and metrics. Learned Publishing, 26(1), 11-17.

Ananny, M., & Crawford, K. (in press). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, doi:10.1177/1461444816676645.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in

Neural Information Processing Systems 29 (NIPS2016) (pp. 4349-4357). Barcelona, Spain: Neural Information Processing Systems Foundation, Inc.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1301-1309). Association for Computational Linguistics.

Chiroro, P. M., Tredoux, C. G., Radaelli, S., & Meissner, C. A. (2008). Recognizing faces across continents: The effect of within-race variations on the own-race bias in face recognition. Psychonomic Bulletin & Review, 15(6), 1089-1092.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism, 3(3), 398-415.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56-62.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies, 2015(1), 92-112.

de O Melo, C., & de Sousa, T. C. (2017). Reflections on cyberethics education for millennial software engineers. In Proceedings of the 1st International Workshop on Software Engineering Curricula for Millennials (pp. 40-46). Los Alamitos, CA: IEEE Press.

Guerini, M., Gatti, L., & Turchi, M. (2013). Sentiment analysis: How to derive prior polarities from SentiWordNet. I Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013) (pp. 1259-1269) https://arxiv.org/abs/1309.5843

Hadžidedić Baždarević, S., & Cristea, A. I. (2017). Do personalisation and emotions affect the use of cancer-related websites? Online Information Review, 41(1), 102-118.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2125-2126). New York, NY: ACM Press.

Halavais, A. (2013). Search engine society. New York, NY: John Wiley & Sons.

Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. In: Proceedings of the Seventeenth International Conference on Machine Learning (ICML00). San Francisco, CA: Morgan Kaufmann Publishers Inc. (pp. 359-366).

Han, H., Otto, C., Liu, X., & Jain, A. K. (2015). Demographic estimation from face images: Human vs. machine performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(6), 1148-1161.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. The ANNALS of the American Academy of Political and Social Science, 659(1), 63-76.

Herman, E. S., & Chomsky, N. (2010). Manufacturing consent: The political economy of the mass media. London, UK: Random House.

Jordan, K. (2015). Exploring the ResearchGate score as an academic metric: reflections and implications for practice. Quantifying and Analysing Scholarly Communication on the Web (ASCW'15). http://ascw.know-center.tugraz.at/wp-content/uploads/2015/06/ASCW15_jordan_response_kraker-lex.pdf.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 3819-3828). New York, NY: ACM Press.

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security, 7(6), 1789-1801.

Lamont, J. (2014). Managing marketing: putting the puzzle together. KM World, 23(10), 12-13.

Li, J., Yang, H., & Zong, C. (2016). Sentiment classification of social media text considering user attributes. In International Conference on Computer Processing of Oriental Languages. Berlin, Germany: Springer International Publishing (pp. 583-594).

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

Liu, J., & Adie, E. (2013). Five challenges in altmetrics: A toolmaker's perspective. Bulletin of the Association for Information Science and Technology, 39(4), 31-34.

Liyakasa, K. (2012). Transforming Into a social CRM enterprise, Customer Relationship Management, 16(6), 38-42.

Mihalcea, R., & Garimella, A. (2016). What men say, what women hear: Finding gender-specific meaning shades. IEEE Intelligent Systems, 31(4), 62-67.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1-21. doi:10.1177/2053951716679679

Montero, C. S., Munezero, M., & Kakkonen, T. (2014). Investigating the role of emotion-based features in author gender classification of text. In International Conference on Intelligent Text Processing and Computational Linguistics. Berlin, Germany: Springer (pp. 98-114).

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes, 45(3), 211-236.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. In Proceedings of the Third International ICWSM Conference (ICWSM2009) Menlo Park, CA: IEEE Press (pp. 278-281).

Paltoglou, G. & Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 1386-1395). Association for Computational Linguistics.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271-279). New York, NY: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1-135.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.

Sweeney, L. (2013). Discrimination in online ad delivery. Queue, 11(3), 1-19.

Teevan, J., Dumais, S., & Horvitz, E. (2004). U.S. Patent Application No. 10/958,560.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

Thelwall, M., Wilkinson, D. & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. Journal of the Association for Information Science and Technology, 61(1), 190-199.

Thelwall, M. & Kousha, K. (2015). ResearchGate: Disseminating, communicating, and measuring Scholarship? Journal of the Association for Information Science and Technology, 66(5), 876-889.

Thelwall, M. (in press). Gender Bias in Sentiment Analysis. Online Information Review. https://dl.dropboxusercontent.com/u/14001543/GenderedSentimentAnalysisLexical.docx

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. Colorado Technology Law Journal, 13(1), 203-217.

Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.

Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. Information Processing & Management, 40(4), 693-707.

Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP2013) ACL Press (pp. 1815-1827).

Volkova, S., & Yoram, B. (2015). On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. Cyberpsychology, Behavior, and Social Networking, 18(12), 726-736. doi:10.1089/cyber.2014.0609.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77-84.

Williams, A. E. (2017). Altmetrics: an overview and evaluation. Online Information Review, 41(3), 311-317.

Wood, P. (2016). The Brits behind Trump. The Spectator [London, UK] (Dec 3, 2016), https://www.spectator.co.uk/2016/12/the-british-data-crunchers-who-say-they-helped-donald-trump-to-win/

Wright, O. (2015). Your tweets are now the Government's business. The Independent [London, UK], 05 June 2015, p. 8.

Yang, Z., Kotov, A., Mohan, A., & Lu, S. (2015). Parametric and non-parametric user-aware sentiment topic models. In Proceedings of the 38th International ACM Conference on Research and Development in Information Retrieval (SIGIR2015). New York, NY: ACM Press (pp. 413-422).